Statistical Methods for Evaluating Prediction Uncertainty in Computer Models

Michael D. McKay
Los Alamos National Laboratory, Los Alamos, NM 87545-0600
mdm@lanl.gov

Abstract

Suppose that the prediction *y* from a model *m* is determined by a vector of input variables *x*. The input variables might define initial conditions of a system being modeled as well as parameter values in the rules determining *y* from the initial conditions. We associate the term *input uncertainty* with a lack of knowledge about appropriate precise input values from which to calculate *y*. Therefore, we treat *x* as a random variable with a probability density function that quantifies input uncertainty. The *prediction distribution* is the corresponding probability distribution induced on *y* by way of the model *m*, and characterizes *prediction uncertainty*. The objective of our analysis is to investigate the relationship between the input variables *x* and the prediction distribution. In particular, we try to identify small *important* subsets *x'* of inputs that "drive" prediction uncertainty.

Two common approaches used for investigation are differential sensitivity analysis and methods based on (linear) regression and correlation coefficients. Generally, these approaches are only valid in the neighborhood of a "nominal value" or they require that *y* be approximately linear in *x*. Furthermore, validity of associated importance measures usually requires that the components of *x* be statistically independent.

Variance-based methodology is so called because of the prominent role played by the variance of the prediction distribution. When the methodology does not depend on the functional form of *m*, as in our case, it is said to be *nonparametric*. Our aim is to find a subset *x'* of the input vector *x* that accounts for a significant part of the variance of the prediction distribution.

The prediction variance can be written $V(y) = V(y') + E(L)$ where *y'* is a function of only the subset *x'* and $E(L) = E(y-y')^2$ is the average or expected squared difference between the full model predictor *y(x)* and a restricted predictor *y'(x')*. If, on average, *y'* is close to *y* then *E(L)* will be small and we would say that the subset *x'* is important because it drives the prediction variance. With the *correlation ratio*, $h^2 = V(y')/V(y)$, as a measure of importance, we now have the problem of how to find the subsets.

Selection of an important subset *x'* has two aspects: the size of the subset and its composition. The problem for us is similar to the variable subset selection problem in regression. Therefore, we proceed in that direction and use an Analysis of Variance (ANOVA) decomposition of variance. However, our decomposition is not based on the usual linear model and does not require statistical independence of the components of *x*. The correlation ratio and the ANOVA *multiple correlation coefficient* R^2 emerge as natural measures of importance. It should be realized that (1) the number of possibilities for subsets of input variables might be astronomical, and (2) suitable estimation of *variance components* might require a large number of computer runs.